Project Description: E-commerce Text Classification

Data Description:

- This is the classification based E-commerce text dataset for 4 categories - "Electronics", "Household", "Books" and "Clothing & Accessories", which almost cover 80% of any E-commerce website.
- The dataset is in ".csv" format with two columns - the first column is the class name and the second one is the data point of that class. The data point is the product and description from the e-commerce website.

Dataset:

The dataset has the following features:

- Data Set Characteristics: Multivariate
- Number of Instances: 50424
- Number of classes: 4

Objective:

To implement the techniques learnt as a part of the course.

Learning Outcomes:

- Basic understanding of text pre-processing.
- What to do after text pre-processing:
  - Bag of words
  - Tf-idf
- Build the classification model.
- Evaluate the Model.

Steps and tasks:

1. Import the libraries, load dataset. (3 Marks)
2. Exploratory Data Analysis and Understanding of data-columns: (12 Marks)
   a. Print Shape of data.
   b. Print data description and info about the data. Comment about the result.
   c. Check the data-type of Text column's first value.
   d. Check for null values and remove the rows in which null values are present.
   e. Check for unique labels in the 'Label' column.
   f. Save the unique labels in the list named 'labels'.
   g. Print first 5 rows of data.
3. Text pre-processing: Data preparation. (15Marks)
   a. Html tag removal.
   b. Remove the numbers.
   c. Tokenization.
   d. Removal of Special Characters and Punctuations.
   e. Conversion to lowercase.
   f. Lemmatize or stemming.
   g. Join the words in the list to convert back to text string in the dataframe. (So that each row contains the data in text format.)
   h. Print first 5 rows of data after pre-processing.
4. Vectorization: (10Marks)
   a. Use CountVectorizer. (use parameter: max_features=1000)
   b. Use TfidfVectorizer. (use parameter: max_features=1000)
5. Fit and evaluate model using both type of vectorization. Print confusion matrix. (6+6Marks)
6. Summarize your understanding of the application of Various Pre-processing and Vectorization and performance of your model on this dataset. (8 Marks)

Happy Learning!

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.